

Ordinary Differential Equation Methods For Markov Decision Processes and Application to Kullback–Leibler Control Cost*

Ana Bušić[†]Sean Meyn[‡]

October 25, 2016

Abstract

A new approach to computation of optimal policies for MDP (Markov decision process) models is introduced. The main idea is to solve not one, but an entire family of MDPs, parameterized by a scalar ζ that appears in the one-step reward function. For an MDP with d states, the family of value functions $\{h_\zeta^* : \zeta \in \mathbb{R}\}$ is the solution to an ODE,

$$\frac{d}{d\zeta} h_\zeta^* = \mathcal{V}(h_\zeta^*)$$

where the vector field $\mathcal{V}: \mathbb{R}^d \rightarrow \mathbb{R}^d$ has a simple form, based on a matrix inverse.

This general methodology is applied to a family of average-cost optimal control models in which the one-step reward function is defined by Kullback-Leibler divergence. The motivation for this reward function in prior work is computation: The solution to the MDP can be expressed in terms of the Perron-Frobenius eigenvector for an associated positive matrix. The drawback with this approach is that no hard constraints on the control are permitted. It is shown here that it is possible to extend this framework to model randomness from nature that cannot be modified by the controller. Perron-Frobenius theory is no longer applicable – the resulting dynamic programming equations appear as complex as a completely unstructured MDP model. Despite this apparent complexity, it is shown that this class of MDPs admits a solution via this ODE technique. This approach is new and practical even for the simpler problem in which randomness from nature is absent.

Keywords: Markov decision processes, Computational methods, Distributed control.

AMS Subject Classifications: 90C40, 93E20, 60J22, 93E35, 60J20, 90C46

1 Introduction

This paper concerns average-cost optimal control for Markovian models. It is assumed that there is a one-step reward w that is a function of state-input pairs. For a given policy that defines the input

*Research supported by French National Research Agency grant ANR-12-MONU-0019, and NSF grants 1609131 and CPS-1259040.

[†]Inria and the Computer Science Department of École Normale Supérieure, Paris, France (ana.busic@inria.fr, <http://www.di.ens.fr/~busic/>).

[‡]Department of Electrical and Computer Engineering at the University of Florida, Gainesville (meyn@ece.ufl.edu, <http://www.meyn.ece.ufl.edu/>).

as a function of present and past state values, the resulting average reward is the limit infimum,

$$\eta = \liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T w(X(t), U(t)) \quad (1)$$

where $\mathbf{X} = \{X(t) : t \geq 0\}$, $\mathbf{U} = \{U(t) : t \geq 0\}$ are the state and input sequences. Under general conditions, the maximum over all policies is deterministic and independent of the initial condition, and the optimal policy is state-feedback — obtained as the solution to the average-reward optimality equations (AROE) [2, 18].

1.1 Background

In this paper the state space \mathbf{X} on which \mathbf{X} evolves is taken to be finite, but possibly large. It is well known that computation of a solution to the AROE may be difficult in such cases. This is one motivation for the introduction of approximation techniques such as reinforcement learning [3, 23].

An interesting alternative is to change the problem so that it is easily solved. In all of the prior work surveyed here, the setting is average-cost optimal control, so that the reward function is replaced by a cost function.

Brockett in [6] introduces a class of controlled Markov models in continuous time. The model and cost function are formulated so that the optimal control problem is easily solved numerically. The ODE technique is applied to this special class of MDPs in Section 2.4.

The theory developed in Section 3 was inspired by the work of Todorov [24], the similar earlier work of Kárný [11], and the more recent work [10, 16]. The state space \mathbf{X} is finite, the action space \mathbf{U} consists of probability mass functions on \mathbf{X} , and the controlled transition matrix is entirely determined by the input as follows:

$$\mathbf{P}\{X(t+1) = x' \mid X(t) = x, U(t) = \mu\} = \mu(x'), \quad x, x' \in \mathbf{X}, \mu \in \mathbf{U}. \quad (2)$$

The MDP has a simple solution only under special conditions on the cost function. It is assumed in [24, 10, 16] that it is the sum two terms: The first is a cost function on \mathbf{X} , which is completely unstructured. The second term is a “control cost”, defined using Kullback–Leibler (K-L) divergence (also known as *relative entropy*).

The control cost is based on deviation from control-free behavior (modeled by a nominal transition matrix P_0). In most applications, P_0 captures randomness from nature. For example, in a queueing model there is uncertainty in inter-arrival times or service times. An optimal solution in this framework would allow modification of arrival statistics and service statistics, which may be entirely infeasible. In this paper the K-L cost framework is broadened to include constraints on the pmf μ appearing in (2).

1.2 Contributions

The new approach to computation proposed in this paper is based on the solution of an entire family of MDP problems. Section 2 begins with a general MDP model in which the one-step reward function is a smooth function of a real parameter $\zeta \in \mathbb{R}$.

For each ζ , the solution to the average-reward optimization problem is based on a relative value function $h_\zeta^* : \mathbf{X} \rightarrow \mathbb{R}$. Under general conditions it is shown that these functions are obtained as the solution to an ordinary differential equation

$$\frac{d}{d\zeta} h_\zeta^* = \mathcal{V}(h_\zeta^*)$$

Consequently, *the solution to an entire family of MDPs can be obtained through the solution of a single ordinary differential equation (ODE).*

Following the presentation of these general results, the paper focuses on the class of MDPs with transition dynamics given in (2): the input space is a subset of the simplex in \mathbb{R}^d , and the cost function c is defined with respect to K-L divergence (see (23) and surrounding discussion). The optimal control formulation is far more general than in the aforementioned work [24, 10, 16], as it allows for inclusion of exogenous randomness in the MDP model.

The dynamic programming equations become significantly more complex in this generality, so that in particular, the Perron-Frobenius computational approach used in prior work is no longer applicable. Nevertheless, the ODE approach can be applied to solve the family of MDP optimal control problems. The vector field $\mathcal{V}: \mathbb{R}^d \rightarrow \mathbb{R}^d$ has special structure that further simplifies computation of the relative value functions.

Simultaneous computation of the optimal policies is essential in applications to “demand dispatch” for providing virtual energy storage from a large collection of flexible loads [1, 16, 7]. In these papers, randomized policies are designed for each of many thousands of electric loads in a distributed control architecture. In this context it is necessary to compute the optimal transition matrix \tilde{P}_ζ for each ζ . Prior to the present work it was not possible to include any exogenous uncertainty in the load model.

In the companion paper [8], the results of the present paper are applied to distributed control of flexible loads, including thermostatically controlled devices such as refrigerators and heating systems. This paper also contains extensions of the ODE method to generate transition matrices with desirable properties, without consideration of optimality.

The remainder of the paper is organized as follows. Section 2 sets the notation for the MDP models in a general setting, and presents an ODE approach to solving the AROE under minimal conditions on the model. Section 3 describes the Kullback–Leibler cost criterion. Special structure of optimal policies obtained in Theorem 3.4 leads to a simple representation of the ODE in Theorem 3.5. Conclusions and topics for future research are contained in Section 4.

2 ODE for MDPs

2.1 MDP model

Consider an MDP with finite state space $\mathbf{X} = \{x^1, \dots, x^d\}$; the action space \mathbf{U} is an open subset of \mathbb{R}^m . The state process is denoted $\mathbf{X} = (X(0), X(1), \dots)$, and the input process $\mathbf{U} = (U(0), U(1), \dots)$. The dynamics of the model are defined by a *controlled transition matrix*: for $x, x' \in \mathbf{X}$, and $u \in \mathbf{U}$, this is defined by

$$P_u(x, x') = \mathbf{P}\{X(t+1) = x' \mid X(t) = x, U(t) = u\}$$

where the right hand side is assumed independent of $t = 0, 1, 2, \dots$.

The one-step reward function is parameterized by a scalar $\zeta \in \mathbb{R}$. It is assumed to be continuously differentiable in this parameter, with derivative denoted

$$W_\zeta(x, u) = \frac{d}{d\zeta} w_\zeta(x, u). \quad (3)$$

Unless there is risk of confusion, dependency on ζ will be suppressed; in particular, we write w rather than w_ζ .

There may be hard constraints: For each $x \in \mathbf{X}$, there is an open set $\mathbf{U}(x) \subset \mathbf{U}$ consisting of feasible inputs $U(t)$ when $X(t) = x$.

The optimal reward η^* is defined to be the maximum of η in (1) over all policies. Under general conditions on the model, η^* is deterministic, and is independent of x . Under further conditions, this value and the optimal policy are characterized by the AROE:

$$\max_{u \in \mathbf{U}(x)} \left\{ w(x, u) + \sum_{x'} P_u(x, x') h^*(x') \right\} = h^*(x) + \eta^* \quad (4)$$

in which the function $h^*: \mathbf{X} \rightarrow \mathbb{R}$ is called the *relative value function*. The stationary policy ϕ^* is obtained from the AROE: $\phi^*(x) \in \mathbf{U}$ is a maximizing value of u in (4) for each x [2, 18].

Structure for the optimal average reward is obtained under minimal assumptions:

Proposition 2.1. *Suppose that the following hold:*

- (i) *The welfare function is affine in its parameter: $w_\zeta(x, u) = w_0(x, u) + \zeta W(x, u)$ for some function W and all x, u .*
- (ii) *For each ζ , the optimal reward η_ζ^* exists, is deterministic, and is independent of the initial condition.*
- (iii) *For each ζ , the optimal reward η_ζ^* is achieved with a stationary policy ϕ_ζ^* , and under this policy, the following ergodic limits exist for each initial condition:*

$$\eta_\zeta^* = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T w_\zeta(X(t), U(t)), \quad \bar{w}_\zeta = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T W(X(t), U(t))$$

Then, η_ζ^* is convex as a function of ζ , with sub-derivative \bar{w}_ζ :

$$\eta_\zeta^* \geq \eta_{\zeta_0}^* + (\zeta - \zeta_0) \bar{w}_{\zeta_0}, \quad \text{for all } \zeta, \zeta_0 \in \mathbb{R}.$$

Proof. Convexity of η_ζ^* will follow from the lower bound. Alternatively, convexity is implied by the linear programming representation of average-cost optimal control, where η_ζ^* is defined as the maximum of linear functions of ζ [15, 4].

To obtain the lower bound, choose any $\zeta, \zeta_0 \in \mathbb{R}$, and consider the average reward based on w_ζ , obtained using $U(t) = \phi_{\zeta_0}^*(X(t))$ for all $t \geq 0$. We then have,

$$\begin{aligned} \eta_\zeta^* &\geq \liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T w_\zeta(X(t), U(t)) \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T w_{\zeta_0}(X(t), U(t)) + \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T (w_\zeta(X(t), U(t)) - w_{\zeta_0}(X(t), U(t))) \end{aligned}$$

The first summation on the right hand side is equal to $\eta_{\zeta_0}^*$. The second reduces to $(\zeta - \zeta_0) \bar{w}_{\zeta_0}$ on substituting $w_\zeta - w_{\zeta_0} = (\zeta - \zeta_0)W$. \square

We next introduce an ODE that solves the AROE for each ζ .

2.2 ODE solution

To construct an ordinary differential equation for h_ζ^* requires several assumptions. The first is a normalization: The relative value function is not unique, since we can add a constant to obtain a new solution. We resolve this by fixing a state $x^\circ \in \mathbf{X}$, and assume that $h_\zeta^*(x^\circ) = 0$ for each ζ .

For any function $h: \mathbf{X} \rightarrow \mathbb{R}$, we define a new function on $\mathbf{X} \times \mathbf{U}$ via,

$$P_u h(x) = \sum_{x'} P_u(x, x') h(x')$$

Similar notation is used for an uncontrolled transition matrix.

Assumptions

(i) For each ζ , a solution to the AROE (h_ζ^*, η_ζ^*) exists, with $h_\zeta(x^\circ) = 0$, and the pair is continuously differentiable in ζ . Moreover, the function of (x, u, ζ) defined by,

$$q_\zeta^*(x, u) = w_\zeta(x, u) + P_u h_\zeta^*(x)$$

is jointly continuously differentiable in (ζ, u) , with the representation

$$\frac{d}{d\zeta} q_\zeta^*(x, u) = W_\zeta(x, u) + P_u H_\zeta^*(x) \quad \text{in which} \quad H_\zeta^*(x) = \frac{d}{d\zeta} h_\zeta^*(x). \quad (5)$$

(ii) The stationary policy exists as the minimum

$$\phi_\zeta^*(x) = \arg \min_{u \in \mathbf{U}(x)} q_\zeta^*(x, u), \quad x \in \mathbf{X},$$

and is continuously differentiable in ζ for each x .

(iii) The optimal transition matrix \tilde{P}_ζ is irreducible, with unique invariant pmf denoted π_ζ , where

$$\tilde{P}_\zeta(x, x') = P_{u^*}(x, x'), \quad u^* = \phi_\zeta^*(x), \quad x, x' \in \mathbf{X}.$$

All of these assumptions hold for the class of MDP models considered in Section 3.

These assumptions imply that for each ζ there is a solution H_ζ to *Poisson's equation*,

$$\check{W}_\zeta + \tilde{P}_\zeta H_\zeta = H_\zeta + \bar{w}_\zeta \quad (6)$$

in which $\check{W}_\zeta(x) = W_\zeta(x, \phi_\zeta^*(x))$, and $\bar{w}_\zeta = \sum_x \pi_\zeta(x) \check{W}_\zeta(x)$. It is assumed throughout that the solution is normalized, with $H_\zeta(x^\circ) = 0$; there is a unique solution to (6) with this normalization [17, Thm. 17.7.2].

The function q_ζ^* is the “Q-function” that appears in Q-learning [3]. Under (i) and (ii), it follows from the AROE that for each x and ζ ,

$$q_\zeta^*(x, \phi_\zeta^*(x)) = \min_{u \in \mathbf{U}(x)} \{w_\zeta(x, u) + P_u h_\zeta^*(x)\} = h_\zeta^*(x) + \eta_\zeta^* \quad (7)$$

MDP vector field

In Theorem 2.2 it is shown that the family of relative value functions solves an ODE. A function $h: \mathbf{X} \rightarrow \mathbb{R}$ is regarded as a vector in \mathbb{R}^d . The vector field \mathcal{V} is not homogeneous, so it is regarded as a mapping $\mathcal{V}: \mathbb{R}^{d+1} \rightarrow \mathbb{R}^d$. For a given function $h: \mathbf{X} \rightarrow \mathbb{R}$ and $\zeta \in \mathbb{R}$, the function $\mathcal{V}(h, \zeta)$ is defined through the following steps:

1. Obtain a policy: $\phi(x) = \arg \max_u \{w_\zeta(x, u) + P_u h(x)\}$.
2. Obtain a transition matrix $\check{P}(x, x') = P_{\phi(x)}(x, x')$, $x, x' \in \mathbf{X}$.
3. Obtain the solution to Poisson's equation, $\check{W} + \check{P}H = H + \bar{w}$, in which $\check{W}(x) = W_\zeta(x, \phi(x))$, $x \in \mathbf{X}$, and \bar{w} is the steady-state mean of \check{W} under this policy. The solution is normalized so that $H(x^\circ) = 0$.
4. Set $\mathcal{V}(h, \zeta) = H$.

Theorem 2.2. *Under the assumptions of this section:*

- (i) *The family of relative value functions $\{h_\zeta^*\}$ is a solution to the ordinary differential equation,*

$$\frac{d}{d\zeta} h_\zeta^* = \mathcal{V}(h_\zeta^*, \zeta) \quad (8)$$

- (ii) $\frac{d}{d\zeta} \eta_\zeta^* = \bar{w}_\zeta$, *with*

$$\bar{w}_\zeta = \sum_x \pi_\zeta(x) W_\zeta(x, \phi_\zeta^*(x)), \quad \text{where } \pi_\zeta \text{ is the invariant pmf for } \check{P}_\zeta.$$

(iii) *If the derivative $W_\zeta(x, u)$ is independent of ζ and u for each x , then the ODE is homogeneous. That is, for each h , the function $\mathcal{V}(h, \zeta)$ does not depend on ζ .*

Proof. The domain of \mathcal{V} is defined to be any h for which the solution to (1)–(3) is possible. The domain may not include all functions h , but it is defined for any of the relative value functions $\{h_\zeta^*\}$; this is true by the assumptions imposed in the theorem.

If W_ζ is independent of ζ and u , then $w_\zeta(x, u) = w_0(x, u) + \zeta W(x)$ for each x, u, ζ . It follows that ϕ is independent of ζ in step 1, and \check{W} is independent of ζ in step 3. Hence the vector field is independent of ζ .

To complete the proof it remains to establish (i), which will lead to the representation for $\frac{d}{d\zeta} \eta_\zeta^*$ in part (ii) of the theorem.

The assumption that \mathbf{U} is open and that q_ζ^* is continuously differentiable is used to apply the first-order condition for optimality of $\phi_\zeta^*(x)$:

$$0 = \frac{\partial}{\partial u} q_\zeta^*(x, u) \Big|_{u=\phi_\zeta^*(x)}$$

On differentiating each side of the AROE in the form (7), we obtain from the chain-rule

$$\begin{aligned} \frac{d}{d\zeta} \{h_\zeta^*(x) + \eta_\zeta^*\} &= \frac{d}{d\zeta} \{q_\zeta^*(x, \phi_\zeta^*(x))\} \\ &= \frac{\partial}{\partial \zeta} q_\zeta^*(x, u) \Big|_{u=\phi_\zeta^*(x)} + \left(\frac{\partial}{\partial u} q_\zeta^*(x, u) \Big|_{u=\phi_\zeta^*(x)} \right) \frac{\partial}{\partial \zeta} \phi_\zeta^*(x) \\ &= \frac{\partial}{\partial \zeta} q_\zeta^*(x, u) \Big|_{u=\phi_\zeta^*(x)} \\ &= \check{W}_\zeta(x) + P_u H_\zeta^*(x) \Big|_{u=\phi_\zeta^*(x)} \end{aligned}$$

where in the last equation we have applied (5). Rearranging terms leads to the fixed point equation

$$\check{W}_\zeta + \check{P}_\zeta H_\zeta^* = H_\zeta^* + \frac{d}{d\zeta} \eta_\zeta^*$$

Taking the mean of each side with respect to π_ζ implies that $\frac{d}{d\zeta} \eta_\zeta^* = \bar{w}_\zeta$. This establishes (ii), and completes the proof that (8) holds. \square

Extensions

An ODE can be constructed for the discounted cost problem with discounted factor $\beta \in (0, 1)$. The DROE (discounted-reward optimality equation) is another fixed point equation, similar to (4):

$$\max_{u \in \mathcal{U}(x)} \left\{ w(x, u) + \beta \sum_{x'} P_u(x, x') h^*(x') \right\} = h^*(x)$$

Step 3 in the construction of the vector field for the ODE is modified as follows: *Obtain the solution to $\check{W} + \beta \check{P}H = H$.* The solution is unique, so no normalization is possible (or needed).

For both average- and discounted-reward settings, an ODE can be constructed when \mathcal{U} is a *finite set* rather than an open subset of \mathbb{R}^m . In this case, under general conditions, the vector field \mathcal{V} is continuous and piecewise smooth, and the optimal policy is piecewise constant as a function of ζ .

We next consider two simple examples to illustrate the conclusions in the average cost setting. The ODE is homogeneous in all of the examples that follow.

2.3 Example 1: Linear-quadratic model

Consider first the simple scalar linear system,

$$X(t+1) = \alpha X(t) + U(t) + N(t+1)$$

in which $0 < \alpha < 1$. The disturbance \mathbf{N} is i.i.d. with zero mean, and finite variance σ_N^2 . The state space and action space are the real line, $\mathbf{X} = \mathcal{U} = \mathbb{R}$. The reward function is taken to be quadratic in the state variable, $w(x, u) = -\zeta x^2 - c(u)$, so that $W(x, u) = -x^2$ is independent of both ζ and u . The cost function $c: \mathbb{R} \rightarrow \mathbb{R}$ is continuously differentiable and convex, its derivative c' is globally Lipschitz continuous, and $c(0) = c'(0) = 0$.

It is assumed in this example that $\zeta \geq 0$. It can be shown that the relative value function h_ζ^* is a concave function of x under these assumptions; it is normalized so that $h_\zeta^*(0) = 0$ for each $\zeta \geq 0$ (that is, $x^\circ = 0$).

The ODE can be developed even in this infinite state-space setting.

Notation is simplified if this is converted to an average-cost optimization problem, with one-step cost function $c_\zeta(x, u) = \zeta x^2 + c(u)$. We let $g_\zeta^* = -h_\zeta^*$, which is a convex function on \mathbb{R} . The AROE becomes the average-cost optimality equation,

$$\min_u \{c_\zeta(x, u) + P_u g_\zeta^*(x)\} = g_\zeta^*(x) + \gamma_\zeta^* \quad (9)$$

with $\gamma_\zeta^* = -\eta_\zeta^*$. The optimal policy is the minimizer,

$$\begin{aligned} \phi_\zeta^*(x) &= \arg \min_u \{c(u) + P_u g_\zeta^*(x)\} \\ &= \arg \min_u \{c(u) + \mathbb{E}[g_\zeta^*(\alpha x + u + N_1)]\} \end{aligned}$$

The ODE is modified as follows. Let \mathcal{K} denote the set of non-negative convex functions $g: \mathbb{R} \rightarrow \mathbb{R}$, and construct the vector field so that $\mathcal{V}: \mathcal{K} \rightarrow \mathcal{K}$. For given $g \in \mathcal{K}$, we must define $\mathcal{V}(g)$. Since we are minimizing cost, step 1 in the construction of the vector field becomes,

Obtain a policy: $\phi(x) = \arg \min_u \{c(u) + \mathbb{E}[g(\alpha x + u + N_1)]\}$.

This is a convex optimization problem whose solution can be obtained numerically.

Step 2 is obtained as follows:

$$\check{P}(x, A) = \mathbb{P}\{\alpha x + \phi(x) + N_1 \in A\}, \quad x \in \mathbf{X}, \quad A \in \mathcal{B}(\mathbf{X}).$$

The solution to Poisson's equation in step 3 of the ODE construction is more challenging. This might be approximated using a basis, as in TD-learning [3, 23].

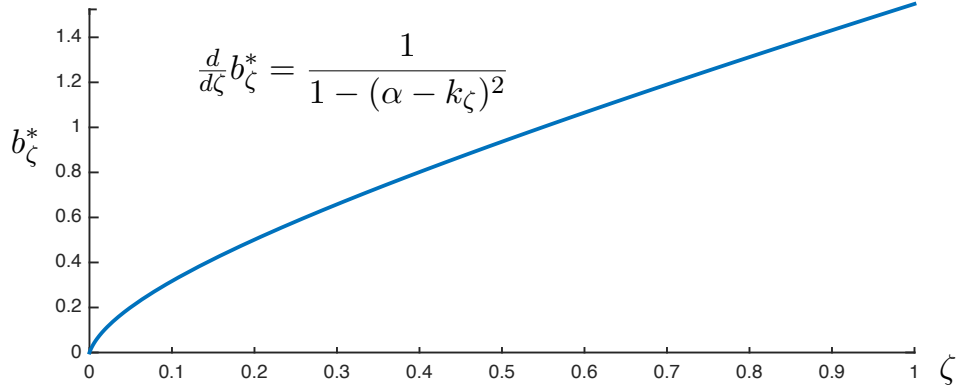


Figure 1: Solution to the ODE (10) for the linear-quadratic example.

If the control cost is quadratic, $c(u) = u^2$, then the relative value function is also quadratic, so that $g_\zeta^*(x) = -h_\zeta^*(x) = bx^2$, with $b \geq 0$, and $b = 0$ only if $\zeta = 0$. The optimal policy is linear, $u = -kx$ for some gain k . The vector field in the ODE can be restricted to functions of this form: For any $b \geq 0$,

(i) Obtain a policy

$$\phi(x) = \arg \min_u \{u^2 + b\mathbf{E}[(\alpha x + u + N_1)^2]\} = \arg \min_u \{u^2 + b[(\alpha x + u)^2 + \sigma_N^2]\}$$

This gives $u = -kx$, with $k = b\alpha/(1 + b)$.

(ii) With a linear policy we obtain, $\tilde{P}(x, A) = \mathbf{P}\{(\alpha - k)x + N_1 \in A\}$.

(iii) Obtain the solution to Poisson's equation, $W + \tilde{P}H = H + \bar{w}$, in which \bar{w} is the steady-state mean of W under this policy. Since $W(x, u) = -x^2$ is quadratic, it follows that H is also quadratic, $H(x) = -Bx^2$, with

$$-B = \frac{1}{1 - (\alpha - k)^2}, \quad \text{and } k \text{ given in (i)}$$

(iv) Set $\mathcal{V}(g) = -H$.

That is, $\frac{d}{d\zeta} g_\zeta^*(x) = Bx^2$, $x \in \mathbb{R}$, $\zeta \geq 0$.

The ODE reduces to a differential equation for the coefficient $b = b_\zeta^*$:

$$\frac{d}{d\zeta} b_\zeta^* = \frac{1}{1 - (\alpha - k_\zeta)^2}, \quad k_\zeta = b_\zeta^* \alpha / (1 + b_\zeta^*)$$

On substitution this simplifies to,

$$\frac{d}{d\zeta} b_\zeta^* = \frac{1}{1 - (\alpha/(1 + b_\zeta^*))^2} \tag{10}$$

with boundary condition $b_\zeta^*|_{\zeta=0} = 0$. Fig. 1 shows the solution to this ODE for $\zeta \in [0, 1]$ with $\alpha = 0.95$.

2.4 Example 2: Brockett's MDP

The theory in this paper is restricted to models in discrete time, but the main results carry over to the continuous time setting. An illustration is provided using a class of MDP models introduced in [6].

As assumed in the present paper, the state space $\mathbf{X} = \{x^1, \dots, x^d\}$ is finite, and the input evolves on an open subset of \mathbb{R}^m . The dynamics are defined by a controlled rate matrix (or generator),

$$A_u(x, x') = A(x, x') + \sum_{k=1}^m u_k B^k(x, x') \quad (11)$$

where A and $\{B^1, \dots, B^m\}$ are $d \times d$ matrices. It is assumed that the input is defined by state feedback $U(t) = \phi(X(t))$. The associated controlled rate matrix

$$\mathcal{A}(x, x') = A_{\phi(x)}(x, x'), \quad x, x' \in \mathbf{X},$$

defines the transition rates for \mathbf{X} under this policy:

$$\mathcal{A}(x, x') = \lim_{t \downarrow 0} \frac{1}{t} [\mathbb{P}\{X(t) = x' \mid X(0) = x\} - \mathbb{I}\{x = x'\}]$$

Adapting the notation to the present paper, the cost function is taken of the form

$$c_\zeta(x, u) = \zeta \kappa(x) + \frac{1}{2} \|u\|^2.$$

in which $\kappa: \mathbf{X} \rightarrow \mathbb{R}$. For the continuous time model, the average-cost optimality equation becomes

$$\min_u \{c_\zeta(x, u) + A_u g_\zeta^*(x)\} = \gamma_\zeta^* \quad (12)$$

in which γ_ζ^* is the optimal average cost, g_ζ^* is the relative value function, and

$$A_u g_\zeta^*(x) = \sum_{x'} A_u(x, x') g_\zeta^*(x'), \quad x \in \mathbf{X}.$$

It is assumed that $g_\zeta^*(x^\circ) = 0$ for some state x° and all ζ .

The minimizer in (12) defines the optimal policy $\phi^*(x)$. For this model and cost function, the minimizer can be obtained by taking the gradient with respect to u and setting this equal to zero to obtain:

$$\phi_k^*(x) = - \sum_{x'} B^k(x, x') g_\zeta^*(x'), \quad x \in \mathbf{X}. \quad (13)$$

The ODE to solve (12) takes the following steps. First, if $\zeta = 0$ then obviously $\phi^* \equiv 0$ and $g_0^* \equiv 0$. This sets the initial condition for the ODE. For other ζ we have as before that $G_\zeta^* = \frac{d}{d\zeta} g_\zeta^*$ solves Poisson's equation for the generator \check{A}_ζ obtained with policy ϕ_ζ^* :

$$\kappa(x) + \sum_j \check{A}_\zeta(x, x') G_\zeta^*(x') = \bar{\kappa}_\zeta, \quad x \in \mathbf{X}, \quad (14)$$

where $\bar{\kappa}_\zeta$ is the steady-state mean of κ under this policy, and

$$\check{A}_\zeta(x, x') = A_{\phi_\zeta^*(x)}(x, x')$$

The following numerical example from [6] will clarify the construction of the ODE. In this example $m = 1$ so that u is scalar-valued, and $\mathbf{X} = \{1, 2, 3\}$. Denote $B = B^1$, where in this example

$$A = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

The input is restricted to $\{u \in \mathbb{R} : u > -1\}$. For $\zeta > 0$, the cost function is designed to penalize the first and third states: $c(1) = c(3) = 3$, and $c(x^2) = 0$. In [6] the case $\zeta = 1/2$ is considered, for which it is shown that $\phi^*(1) = \phi^*(3) = \sqrt{12} - 3$, and $\phi^*(x^2) = 0$.

Written in vector form, Poisson's equation (14) becomes

$$3b + A_\zeta v_\zeta = \bar{\kappa}_\zeta e \quad (15)$$

in which $b^T = (1, 0, 1)$, $e^T = (1, 1, 1)$, $v_\zeta(i) = G_\zeta^*(x^i)$, and

$$A_\zeta(i, j) = A(i, j) + \phi_\zeta^*(x^i)B(i, j), \quad 1 \leq i, j \leq 3.$$

This example is designed to have simple structure. From the form of the optimal policy (13), it follows that $\phi_\zeta^*(2) = 0$ for any ζ . Moreover, from symmetry of the model it can be shown that $\phi_\zeta^*(1) = \phi_\zeta^*(3)$, and $g_\zeta^*(1) = g_\zeta^*(3)$. We take $x^\circ = 2$ so that $g_\zeta^*(2) = 0$. Consequently, the 3-dimensional ODE for g_ζ^* will reduce to a one-dimensional ODE for $\xi_\zeta := g_\zeta^*(1)$.

The expression for the optimal policy (13) also gives

$$\phi_\zeta^*(1) = -\xi_\zeta \sum_j B(1, j)b(j) = \xi_\zeta$$

And, since the second row of B is zero, it follows that $A_\zeta = A + \xi_\zeta B$.

We have $\frac{d}{d\zeta}\xi_\zeta = G_\zeta^*(1) = v_\zeta(1)$, and Poisson's equation (15) becomes

$$3b + v_\zeta(1)A_\zeta b = \bar{\kappa}_\zeta e$$

The first two rows of this vector equation give

$$\begin{aligned} 3 + [-1 + \xi_\zeta(-1)]v_\zeta(1) &= \bar{\kappa}_\zeta \\ 0 + 2v_\zeta(1) &= \bar{\kappa}_\zeta \end{aligned}$$

Substituting the second equation into the first gives

$$\frac{d}{d\zeta}\xi_\zeta = v_\zeta(1) = \frac{3}{3 + \xi_\zeta}$$

On making the change of variables $f_\zeta = 3 + \xi_\zeta$ we obtain

$$\frac{1}{2} \frac{d}{d\zeta} f_\zeta^2 = f_\zeta \frac{d}{d\zeta} f_\zeta = 3,$$

whose solution is given by $f_\zeta^2 = f_0^2 + 6\zeta$, with $f_0^2 = 9$.

In summary, $\phi_\zeta^*(1) = \xi_\zeta = -3 + f_\zeta$, giving

$$\phi_\zeta^*(1) = \phi_\zeta^*(3) = -3 + \sqrt{9 + 6\zeta}, \quad \phi_\zeta^*(2) = 0.$$

It is necessary to restrict ζ to the interval $(-5/6, \infty)$ to respect the constraint that $\phi_\zeta^*(x) > -1$ for all x .

Based on the formula $\frac{d}{d\zeta}\gamma_\zeta^* = \bar{\kappa}_\zeta$ and the preceding formula $\bar{\kappa}_\zeta = 2v_\zeta(1) = 2\frac{d}{d\zeta}\xi_\zeta$, it follows that

$$\gamma_\zeta^* = 2\xi_\zeta = -6 + 2\sqrt{9 + 6\zeta}, \quad \zeta > -5/6;$$

a concave function of ζ , as predicted by Prop. 2.1.

3 MDPs with Kullback–Leibler Cost

The general results of the previous section are now applied to a particular class of MDP models.

3.1 Assumptions and Notation

The dynamics of the MDP are assumed of the form (2), where the action space consists of a convex subset of probability mass functions on \mathbf{X} . The welfare function is assumed to be affine in ζ , as assumed in Prop. 2.1. To maintain notational consistency with prior work [7, 16, 8] we denote

$$w_\zeta = w_0 + \zeta \mathcal{U}, \quad \zeta \in \mathbb{R}, \quad (16)$$

and assume that $\mathcal{U}: \mathbf{X} \rightarrow \mathbb{R}$ is a function of the state only. In the notation of Prop. 2.1, we have $W(x, u) = \mathcal{U}(x)$ for all x, u . Under these conditions, it was shown in Theorem 2.2 that the ODE (8) is homogeneous.

The first term w_0 in (16) is the negative of a control cost. Its definition begins with the specification of a transition matrix P_0 that describes nominal (control-free) behavior. It is assumed to be *irreducible and aperiodic*. Equivalently, there is $n_0 \geq 1$ such that for each $x, x' \in \mathbf{X}$,

$$P_0^n(x, x') > 0, \quad x \in \mathbf{X}, \quad n \geq n_0. \quad (17)$$

It follows that P_0 admits a unique invariant pmf, denoted π_0 .

In the MDP model we deviate from this nominal behavior, but restrict to transition matrices satisfying $P(x, \cdot) \prec P_0(x, \cdot)$ for each x . In fact, the optimal solutions will be equivalent:

$$P(x, x') > 0 \iff P_0(x, x') > 0, \quad \text{for all } x, x' \in \mathbf{X} \quad (18)$$

Under this condition it follows that P is also irreducible and aperiodic.

The following representation will be used in different contexts throughout the paper. Any function $h: \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ is regarded as an unnormalized log-likelihood ratio: Denote for $x, x' \in \mathbf{X}$,

$$P_h(x, x') := P_0(x, x') \exp(h(x' | x) - \Lambda_h(x)), \quad (19)$$

in which $h(x' | x)$ is the value of h at $(x, x') \in \mathbf{X} \times \mathbf{X}$, and $\Lambda_h(x)$ is the normalization constant,

$$\Lambda_h(x) := \log \left(\sum_{x'} P_0(x, x') \exp(h(x' | x)) \right) \quad (20)$$

For any transition matrix P , an invariant pmf is interpreted as a row vector, so that invariance can be expressed $\pi P = \pi$. Any function $f: \mathbf{X} \rightarrow \mathbb{R}$ is interpreted as a d -dimensional column vector, and we use the standard notation $Pf(x) = \sum_{x'} P(x, x') f(x')$, $x \in \mathbf{X}$.

The *fundamental matrix* is the inverse,

$$Z = [I - P + 1 \otimes \pi]^{-1} \quad (21)$$

where $1 \otimes \pi$ is a matrix in which each row is identical, and equal to π . If P is irreducible and aperiodic, then it can be expressed as a power series:

$$Z = \sum_{n=0}^{\infty} [P - 1 \otimes \pi]^n \quad (22)$$

with $[P - 1 \otimes \pi]^0 := I$ (the $d \times d$ identity matrix), and $[P - 1 \otimes \pi]^n = P^n - 1 \otimes \pi$ for $n \geq 1$.

The *Donsker-Varadhan rate function* is denoted,

$$K(P\|P_0) = \sum_{x,x'} \pi(x)P(x,x') \log\left(\frac{P(x,x')}{P_0(x,x')}\right) \quad (23)$$

Letting $\Pi(x, x') = \pi(x)P(x, x')$ and $\Pi_0(x, x') = \pi(x)P_0(x, x')$, we have

$$K(P\|P_0) = D(\Pi\|\Pi_0) \quad (24)$$

where D denotes K-L divergence. It is called a “rate function” because it defines the relative entropy rate between two stationary Markov chains, and appears in the theory of large deviations for Markov chains [12].

For the transition matrix P_h defined in (19), the rate function can be expressed in terms of its invariant pmf π_h , the bivariate pmf $\Pi_h(x, x') = \pi_h(x)P_h(x, x')$, and the log moment generating function (20):

$$\begin{aligned} K(P_h\|P_0) &= \sum_{x,x'} \Pi_h(x, x') [h(x' | x) - \Lambda_h(x)] \\ &= \sum_{x,x'} \Pi_h(x, x') h(x' | x) - \sum_x \pi_h(x) \Lambda_h(x) \end{aligned} \quad (25)$$

As in [24, 10, 16], the rate function is used here to model the cost of deviation from the nominal transition matrix P_0 : the control objective in this prior work can be cast as the solution to the convex optimization problem,

$$\eta_\zeta^* = \max_{\pi, P} \{ \zeta \pi(\mathcal{U}) - K(P\|P_0) : \pi P = \pi \} \quad (26)$$

where $\mathcal{U}: \mathbf{X} \rightarrow \mathbb{R}$, and the maximum is over all transition matrices.

Nature & nurture In many applications it is necessary to include a model of randomness from nature along with the randomness introduced by the local control algorithm (nurture). This imposes additional constraints in the optimization problem (26).

Consider a Markov model in which the full state space is the cartesian product of two finite state spaces: $\mathbf{X} = \mathbf{X}_u \times \mathbf{X}_n$, where \mathbf{X}_u are components of the state that can be directly manipulated through control. The “nature” components \mathbf{X}_n are not subject to direct control. For example, these variables may be used to model service and arrival statistics in a queueing model, or uncertainty in terrain in an application to robotics.

Elements of \mathbf{X} are denoted $x = (x_u, x_n)$. Any state transition matrix under consideration is assumed to have the following conditional-independence structure,

$$P(x, x') = R(x, x'_u)Q_0(x, x'_n), \quad x \in \mathbf{X}, x'_u \in \mathbf{X}_u, x'_n \in \mathbf{X}_n \quad (27)$$

where $\sum_{x'_u} R(x, x'_u) = \sum_{x'_n} Q_0(x, x'_n) = 1$ for each x . The matrix Q_0 is out of our control – this models dynamics such as the weather.

To underscore the generality of this model, consider a standard MDP model with finite state space \mathbf{S} , finite action space \mathbf{A} , and controlled transition law ϱ . Letting Φ denote the state process and U the input process, we have for any two states s, s' , and any action a ,

$$\mathbb{P}\{\Phi(t+1) = s' \mid \Phi(t) = s, U(t) = a\} = \varrho(s' \mid s, a)$$

A randomized policy is defined by a function $\phi: \mathbf{A} \times \mathbf{S} \rightarrow [0, 1]$ for which $\phi(\cdot \mid s)$ is a probability law on \mathbf{A} for each $s \in \mathbf{S}$.

Proposition 3.1. *Consider the MDP model with transition law ϱ and randomized policy ϕ . For each $t \geq 0$ denote $X_n(t) = \Phi(t)$ and $X_u(t) = U(t-1)$, where $X(0) = (U(-1), \Phi(0))$ is the initial condition. Then $\mathbf{X} = (\mathbf{X}_u, \mathbf{X}_n)$ is a Markov chain on $\mathbf{X} = \mathbf{A} \times \mathbf{S}$, with transition matrix of the form (27), where for $x, x' \in \mathbf{X}$,*

$$Q_0(x, x'_n) = \varrho(x'_n | x_n, x_u), \quad R(x, x'_u) = \phi(x'_u | x_n),$$

Proof. From the definitions and Bayes' rule,

$$\begin{aligned} & \mathbb{P}\{X_u(t+1) = x'_u, X_n(t+1) = x'_n | X(t) = x\} \\ &= \mathbb{P}\{X_n(t+1) = x'_n | X_u(t+1) = x'_u, X(t) = x\} \mathbb{P}\{X_u(t+1) = x'_u | X(t) = x\} \\ &= \mathbb{P}\{\Phi(t+1) = x'_n | U(t) = x'_u, X(t) = x\} \mathbb{P}\{U(t) = x'_u | X(t) = x\} \end{aligned}$$

Recall that $X(t) = (\Phi(t), U(t-1))$. The pair $(U(t-1), \Phi(t+1))$ are conditionally independent given $(\Phi(t), U(t))$, so that the right hand side becomes,

$$\mathbb{P}\{\Phi(t+1) = x'_n | U(t) = x'_u, \Phi(t) = x_n\} \mathbb{P}\{U(t) = x'_u | \Phi(t) = x_n\}$$

This establishes the desired result:

$$\begin{aligned} & \mathbb{P}\{X_u(t+1) = x'_u, X_n(t+1) = x'_n | X(t) = x\} \\ &= \varrho(x'_n | x_n, x_u) \phi(x'_u | x_n) \end{aligned}$$

□

3.2 Optimal control with Kullback–Leibler cost

We consider now the optimization problem (26), subject to the structural constraint (27), with Q_0 fixed. The maximizer defines a transition matrix that is denoted,

$$\check{P}_\zeta = \arg \max_P \{ \zeta \pi(\mathcal{U}) - K(P \| P_0) : \pi P = \pi \} \quad (28)$$

It is shown in Prop. 3.3 that this can be cast as a convex program, even when subject to the structural constraint (27). The optimization variable in the convex program will be taken to be pmfs Π on the product space $\mathbf{X} \times \mathbf{X}_u$.

Define for each π and R the pmf on $\mathbf{X} \times \mathbf{X}_u$,

$$\Pi_{\pi, R}(x, x'_u) = \pi(x) R(x, x'_u), \quad x \in \mathbf{X}, x'_u \in \mathbf{X}_u. \quad (29)$$

The pmf π can be recovered from $\Pi_{\pi, R}$ via $\pi(x) = \sum_{x'_u} \Pi_{\pi, R}(x, x'_u)$, $x \in \mathbf{X}$, and the matrix R can also be recovered via $R(x, x'_u) = \Pi_{\pi, R}(x, x'_u) / \pi(x)$, provided $\pi(x) > 0$.

The following result shows that we can restrict to R for which $\Pi_{\pi, R} \prec \Pi_{\pi_0, R_0}$.

Lemma 3.2. *For any transition matrix P ,*

$$K(P \| P_0) < \infty \iff \Pi_{\pi, R} \prec \Pi_{\pi_0, R_0}$$

Proof. If $K(P \| P_0) < \infty$ then $P(x, \cdot) \prec P_0(x, \cdot)$. This implies that $R(x, \cdot) \prec R_0(x, \cdot)$ for each $x \in \mathbf{X}$ satisfying $\pi(x) > 0$, and also that $\pi \prec \pi_0$.

Hence, if $K(P\|P_0) < \infty$, then for each x and x'_u ,

$$\begin{aligned}\pi_0(x)R_0(x, x'_u) = 0 &\Rightarrow \pi(x)R_0(x, x'_u) = 0 \\ &\Rightarrow \pi(x)R(x, x'_u) = 0,\end{aligned}$$

which establishes one implication: $\Pi_{\pi, R} \prec \Pi_{\pi_0, R_0}$ whenever $K(P\|P_0) < \infty$.

Conversely, if $\Pi_{\pi, R} \prec \Pi_{\pi_0, R_0}$ then $K(P\|P_0) < \infty$ by the definition of K , and the convention $s \log(s) = 0$ when $s = 0$. \square

Lemma 3.2 is one step towards the proof of the following convex program representation of (26):

Proposition 3.3. *The objective function in (26) is concave in the variable $\Pi = \Pi_{\pi, R}$, subject to the convex constraints,*

$$\Pi \text{ is a pmf on } \mathsf{X} \times \mathsf{X}_u \tag{30a}$$

$$\Pi \prec \Pi_0, \quad \text{with } \Pi_0(x, x'_u) = \pi_0(x)R_0(x, x'_u) \tag{30b}$$

$$\sum_x Q_0(x, x'_n)\Pi(x, x'_u) = \sum_{x_u} \Pi(x', x_u) \quad \text{for } x' = (x'_u, x'_n) \in \mathsf{X} \tag{30c}$$

It admits an optimizer $\Pi^*(x, x'_u) = \tilde{\pi}_\zeta(x)\tilde{R}_\zeta(x, x'_u)$, in which $\tilde{\pi}_\zeta(x) > 0$ for each x . Consequently, there exists an optimizer \tilde{P}_ζ for (28), with invariant pmf $\tilde{\pi}_\zeta$.

Proof. We first consider the constraints: (a) is by definition, and (c) is the invariance constraint for (π, P) . Constraint (b) is without loss of generality, given Lemma 3.2.

Next we turn to the objective function: The function to be maximized in (26) can be expressed

$$\zeta\pi(\mathcal{U}) - K(P\|P_0) = \sum_x \pi(x)w(x, R)$$

in which

$$\begin{aligned}w(x, R) &= \zeta\mathcal{U}(x) - \sum_{x'} P(x, x') \log\left(\frac{P(x, x')}{P_0(x, x')}\right) \\ &= \zeta\mathcal{U}(x) - \sum_{x'_u} R(x, x'_u) \log\left(\frac{R(x, x'_u)}{R_0(x, x'_u)}\right)\end{aligned} \tag{31}$$

The second equation follows from the assumption that P depends on R through (27). Multiplying each side by $\pi(x)$ and summing over x we obtain a representation in terms of the variable $\Pi_{\pi, R}$, with $\zeta\pi(\mathcal{U}) - K(P\|P_0) =$

$$\zeta \sum_{x, x'_u} \Pi_{\pi, R}(x, x'_u) \mathcal{U}(x) - D(\Pi_{\pi, R} \| \Pi_{\pi, R_0})$$

The K-L divergence D is known to be jointly convex in its two arguments [9]. Since Π_{π, R_0} is a linear function of $\Pi_{\pi, R}$, this establishes concavity.

The existence of an optimizer follows from the fact that the function to be optimized is continuous as a function of $\Pi_{\pi, R}$, and the domain of optimization (30a–30c) is compact. \square

It is shown in Theorem 3.4 that the optimal value η_ζ^* together with a *relative value function* h_ζ^* solve the average reward optimization equation (AROE):

$$\max_R \left\{ w(x, R) + \sum_{x'} P(x, x') h_\zeta^*(x') \right\} = h_\zeta^*(x) + \eta_\zeta^* \quad (32)$$

Recall that the relative value function is not unique, since a new solution is obtained by adding a non-zero constant; the normalization $h_\zeta^*(x^\circ) = 0$ is imposed, where $x^\circ \in \mathbf{X}$ is a fixed state.

The proof of Theorem 3.4 is given in the Appendix.

Theorem 3.4. *There exist optimizers $\{\tilde{\pi}_\zeta, \tilde{P}_\zeta : \zeta \in \mathbb{R}\}$, and solutions to the AROE $\{h_\zeta^*, \eta_\zeta^* : \zeta \in \mathbb{R}\}$ with the following properties:*

(i) *The optimizer \tilde{P}_ζ can be obtained from the relative value function h_ζ^* as follows:*

$$\tilde{P}_\zeta(x, x') := P_0(x, x') \exp(h_\zeta(x'_u | x) - \Lambda_{h_\zeta}(x)) \quad (33)$$

where for $x \in \mathbf{X}$, $x'_u \in \mathbf{X}_u$,

$$h_\zeta(x'_u | x) = \sum_{x'_n} Q_0(x, x'_n) h_\zeta^*(x'_u, x'_n), \quad (34)$$

and $\Lambda_{h_\zeta}(x)$ is the normalizing constant (20) with $h = h_\zeta$.

(ii) $\{\tilde{\pi}_\zeta, \tilde{P}_\zeta, h_\zeta^*, \eta_\zeta^* : \zeta \in \mathbb{R}\}$ are continuously differentiable in the parameter ζ .

□

The fact that the domain of optimization (30a–30c) is compact was helpful in establishing the existence of an optimizer. However, the results in Section 2 require that the action space be an *open* set. To apply the results of Section 2 we can apply Theorem 3.4 (i), which justifies the restriction (18). The restricted action space is an open subset of \mathbb{R}^m for some $m < d$.

Representations for the derivatives in Theorem 3.4 (ii), in particular the derivative of $\Lambda_{h_\zeta^*}$ with respect to ζ , lead to a representation for the ODE used to compute the optimal transition matrices $\{\tilde{P}_\zeta\}$.

3.3 ODE Solution

It is shown here that the assumptions of Theorem 2.2 hold, and hence the relative value functions $\{h_\zeta^* : \zeta \in \mathbb{R}\}$ can be obtained as the solution to an ODE.

At the start of Section 2 it is assumed that the action space is an open subset of Euclidean space, and this assumption is required in Theorem 2.2. This can be imposed without loss of generality since any optimizer satisfies (18).

It is convenient to generalize the problem slightly here. Let $\{h_\zeta^\circ : \zeta \in \mathbb{R}\}$ denote a family of functions on \mathbf{X} , continuously differentiable in the parameter ζ . They are not necessarily relative value functions, but we maintain the structure established in Theorem 3.4 for the family of transition matrices. Denote,

$$h_\zeta(x'_u | x) = \sum_{x'_n} Q_0(x, x'_n) h_\zeta^\circ(x'_u, x'_n), \quad x \in \mathbf{X}, x'_u \in \mathbf{X}_u \quad (35)$$

and then define as in (19),

$$P_\zeta(x, x') := P_0(x, x') \exp(h_\zeta(x'_u | x) - \Lambda_{h_\zeta}(x)) \quad (36)$$

The function $\Lambda_{h_\zeta} : \mathsf{X} \rightarrow \mathbb{R}$ is a normalizing constant, exactly as in (20):

$$\Lambda_{h_\zeta}(x) := \log \left(\sum_{x'} P_0(x, x') \exp(h_\zeta(x'_u | x)) \right)$$

We begin with a general method to construct a family of functions $\{h_\zeta^\circ : \zeta \in \mathbb{R}\}$ based on an ODE. Using notation similar to Theorem 2.2, the ODE is expressed,

$$\frac{d}{d\zeta} h_\zeta^\circ = \mathcal{V}(h_\zeta^\circ), \quad \zeta \in \mathbb{R}, \quad (37)$$

with boundary condition $h_0^\circ \equiv 0$. A particular instance of the method will result in $h_\zeta^\circ = h_\zeta^*$ for each ζ .

Assumed given is a mapping \mathcal{H}° from transition matrices to functions on X . Following this, the vector field \mathcal{V} is obtained through the following two steps: For a function $h : \mathsf{X} \rightarrow \mathbb{R}$,

- (i) Define a new transition matrix via (19),

$$P_h(x, x') := P_0(x, x') \exp(h(x'_u | x) - \Lambda_h(x)), \quad x, x' \in \mathsf{X}, \quad (38)$$

in which $h(x'_u | x) = \sum_{x'_n} Q_0(x, x'_n) h(x'_u, x'_n)$, and $\Lambda_h(x)$ is a normalizing constant.

- (ii) Compute $H^\circ = \mathcal{H}^\circ(P_h)$, and define $\mathcal{V}(h) = H^\circ$. It is assumed that the functional \mathcal{H}° is constructed so that $H^\circ(x^\circ) = 0$ for any P .

In [8] the functional \mathcal{H}° is designed to ensure desirable properties in the “demand dispatch” application that is the focus of that paper. It is shown here that a particular choice of the function \mathcal{H}° will provide the solution to the collection of MDPs (26). Its domain will include only transition matrices that are irreducible and aperiodic. For any transition matrix P in this domain, the fundamental matrix Z is obtained using (21), and then $H^\circ = \mathcal{H}^\circ(P)$ is defined as

$$H^\circ(x) = \sum_{x'} [Z(x, x') - Z(x^\circ, x')] \mathcal{U}(x'), \quad x \in \mathsf{X} \quad (39)$$

The function H° is a solution to Poisson’s equation,

$$PH^\circ = H^\circ - \mathcal{U} + \bar{u} \quad (40)$$

where \bar{u} (also written $\pi(\mathcal{U})$) is the steady-state mean of \mathcal{U} :

$$\bar{u} := \sum_x \pi(x) \mathcal{U}(x) \quad (41)$$

The proof of Theorem 3.5 is given in the Appendix.

Theorem 3.5. *Consider the ODE (37) with boundary condition $h_0^\circ \equiv 0$, and with $H^\circ = \mathcal{H}^\circ(P)$ defined using (39) for each transition matrix P that is irreducible and aperiodic.*

The solution to this ODE exists, and the resulting functions $\{h_\zeta^\circ : \zeta \in \mathbb{R}\}$ coincide with the relative value functions $\{h_\zeta^ : \zeta \in \mathbb{R}\}$. Consequently, $\tilde{P}_\zeta = P_{h_\zeta}$ for each ζ . \square*

We sketch here the main ideas of the proof of Theorem 3.5.

The Implicit Function Theorem is used to establish differentiability of the relative value functions and average reward as a function of ζ . The ODE representation can then be obtained from Theorem 2.2.

The next step is to establish the particular form for the ODE. The statement of the theorem is equivalent to the representation $H_\zeta^* = \mathcal{H}^\circ(\check{P}_\zeta)$ for each ζ , where h_ζ^* is the relative value function, \check{P}_ζ is defined in (28), and

$$H_\zeta^* = \frac{d}{d\zeta} h_\zeta^* \quad (42)$$

The first step in the proof of (42) is a fixed point equation that follows from the AROE. The following identity is given in Prop. A.2:

$$\zeta \mathcal{U} + \Lambda_{h_\zeta^*} = h_\zeta^* + \eta_\zeta^*. \quad (43)$$

A representation for the derivative of the log moment generating function is obtained in Lemma B.4,

$$\frac{d}{d\zeta} \Lambda_{h_\zeta^*}(x) = \sum_{x'} \check{P}_\zeta(x, x') H_\zeta^*(x').$$

Differentiating each side of (43) then gives,

$$\mathcal{U} + \check{P}_\zeta H_\zeta^* = H_\zeta^* + \frac{d}{d\zeta} \eta_\zeta^*. \quad (44)$$

This is Poisson's equation, and it follows that $\tilde{\pi}_\zeta(\mathcal{U}) = \frac{d}{d\zeta} \eta_\zeta^*$. Moreover, since $h_\zeta^*(x^\circ) = 0$ for every ζ , we must have $H_\zeta^*(x^\circ) = 0$ as well. Since the solution to Poisson's equation with this normalization is unique, we conclude that (42) holds, and hence $H_\zeta^* = \mathcal{H}^\circ(\check{P}_\zeta)$ as claimed.

4 Conclusions

It is surprising that an MDP can be solved using an ODE under general conditions, and fortunate that this ODE admits simple structure in the K-L cost framework that is a focus of the paper.

It is likely that the ODE has special structure for other classes of MDPs, such as the “rational inattention” framework of [21, 22, 19, 20]. The computational efficiency of this approach will depend in part on numerical properties of the ODE, such as its sensitivity for high dimensional models.

Finally, it is hoped that this approach will lead to new approaches to approximate dynamic programming or reinforcement learning.

Appendices

A AROE and Duality

Based on the linear programming (LP) approach to dynamic programming [15, 4], it can be shown that the AROE is the dual of the primal problem (26). The relative value function h^* is the dual variable associated with the invariance constraint $\pi = \pi P$ [4]. To prove Theorem 3.4 we require properties of the primal and dual.

The primal (26) is equivalently expressed,

$$\eta_\zeta^* = \max_{\pi, R} \left\{ \sum_x \pi(x) w(x, R) : (27) \text{ holds, and } \pi = \pi P \right\} \quad (45)$$

The AROE becomes,

$$\max_R \left\{ w(x, R) + \sum_{x'} P(x, x') h_\zeta^*(x') \right\} = h_\zeta^*(x) + \eta_\zeta^* \quad (46)$$

It will be shown that (46) can be interpreted as a dual of the convex program (45). We first characterize its optimizer, denoted \check{R}_ζ . This representation is based on the convex duality between K-L divergence and the log-moment generating function recalled in Lemma A.1.

Fix a pmf μ_0 on X . For any function $F: \mathsf{X} \rightarrow \mathbb{R}$, the log-moment generating function is denoted

$$\Lambda(F) = \log \left\{ \sum_x \mu_0(x) \exp(F(x)) \right\}$$

The mean of a function F under an arbitrary pmf μ is denoted $\mu(F) = \sum_x \mu(x) F(x)$. The following lemma can be regarded as a consequence of Kullback's inequality (see eqn (4.5) of [13]); see also Theorem 3.1.2 of [9].

Lemma A.1. *The log-moment generating function is the convex dual of relative entropy,*

$$\Lambda(F) = \max_{\mu} \{ \mu(F) - D(\mu \| \mu_0) \}$$

where the maximum is over all pmfs, and is achieved uniquely with,

$$\mu^*(x) = \mu_0(x) \exp\{F(x) - \Lambda(F)\}, \quad x \in \mathsf{X}.$$

□

The following representation for \check{R}_ζ easily follows. The fixed point equation appearing in Prop. A.2 was previously stated in (43).

Proposition A.2. *The control matrix maximizing the left hand side of (46) is given by,*

$$\check{R}_\zeta(x, x'_u) = R_0(x, x'_u) \exp(h_\zeta^*(x'_u | x) - \Lambda_{h_\zeta^*}(x)). \quad (47)$$

Consequently, the AROE is equivalent to the fixed point equation $\zeta \mathcal{U} + \Lambda_{h_\zeta^*} = h_\zeta^* + \eta_\zeta^*$.

Proof. Using (31), the AROE becomes $h_\zeta^*(x) + \eta^* =$

$$\max_R \left\{ w(x, R) + \sum_{x'_u, x'_n} R(x, x'_u) Q_0(x, x'_n) h_\zeta^*(x'_u, x'_n) \right\}.$$

Recalling the notation $h_\zeta^*(x'_u | x)$ in (34), we obtain $h_\zeta^*(x) + \eta^* =$

$$\zeta \mathcal{U}(x) + \max_R \left\{ \sum_{x'_u} R(x, x'_u) h_\zeta^*(x'_u | x) - \sum_{x'_u} R(x, x'_u) \log \left(\frac{R(x, x'_u)}{R_0(x, x'_u)} \right) \right\} \quad (48)$$

For fixed x denote $F(x'_u) = h_\zeta^*(x'_u | x)$, $\mu_0(x'_u) = R_0(x, x'_u)$ and $\mu(x'_u) = R(x, x'_u)$, $x'_u \in \mathsf{X}_u$. The maximization variable in (48) is μ , and the maximization problem we must solve is,

$$\max_{\mu} \{ \mu(F) - D(\mu \| \mu_0) \}$$

The formula for the optimizer μ^* in Lemma A.1 gives the expression for \check{R}_ζ in (47).

The fact that the optimal value is $\Lambda(F)$ implies the fixed point equation (43). □

It is established next that the AROE does indeed hold, by constructing a dual of (45) obtained through a relaxation of the invariance constraint. A dual functional φ_ζ^* is defined for any function $h: \mathsf{X} \rightarrow \mathbb{R}$ via

$$\varphi_\zeta^*(h) = \max_{\pi, R} \left\{ \sum_x \pi(x) [w(x, R) + (P - I)h(x)] \right\}$$

where (π, R) are now independent variables, and P is obtained from R via (27). We have $\varphi_\zeta^*(h) \geq \eta_\zeta^*$ for any h , and there is no duality gap:

Proposition A.3. *There exists h_ζ^* such that $\varphi_\zeta^*(h_\zeta^*) = \eta_\zeta^*$. The pair $(h_\zeta^*, \eta_\zeta^*)$ is a solution to the AROE (32).*

Proof. To show that there is no duality gap we apply Prop. 3.3, which establishes that the primal is a convex program, and hence a sufficient condition is Slater's condition [5, Section 5.3.2]. This condition holds because Π_{π_0, R_0} is in the relative interior of the constraint-set for the primal.

Since there is no duality gap, it then follows that there exists a maximizer for φ_ζ^* , denoted h_ζ^* , which satisfies $\eta_\zeta^* = \varphi_\zeta^*(h_\zeta^*)$. To obtain the AROE, consider this representation: $\eta_\zeta^* =$

$$\max_{\pi} \left\{ \sum_x \pi(x) \max_R [w(x, R) + Ph_\zeta^*(x) - h_\zeta^*(x)] \right\}$$

The maximum over pmfs π is the same as the maximum over x :

$$\eta_\zeta^* = \max_x \max_R \{w(x, R) + Ph_\zeta^*(x) - h_\zeta^*(x)\}$$

To complete the proof we must remove the maximum over x . For this, recall that π_0 and hence $\tilde{\pi}_\zeta$ have full support (they are strictly positive on all of X).

Prop. A.2 implies that the maximum over R is uniquely given by \check{R}_ζ in (47), so that

$$\eta_\zeta^* = \max_x \{w(x, \check{R}_\zeta) + \check{P}_\zeta h_\zeta^*(x) - h_\zeta^*(x)\}$$

Averaging over the optimizing pmf $\tilde{\pi}_\zeta$ gives, by invariance,

$$\begin{aligned} \eta_\zeta^* &= \sum_x \tilde{\pi}_\zeta(x) w(x, \check{R}_\zeta) \\ &= \sum_x \tilde{\pi}_\zeta(x) \{w(x, \check{R}_\zeta) + \check{P}_\zeta h_\zeta^*(x) - h_\zeta^*(x)\}. \end{aligned}$$

Because $\tilde{\pi}_\zeta(x) > 0$ for every x , it follows that the AROE (46) holds:

$$\begin{aligned} \eta_\zeta^* &= w(x, \check{R}_\zeta) + \check{P}_\zeta h_\zeta^*(x) - h_\zeta^*(x) \\ &= \max_R \{w(x, R) + Ph_\zeta^*(x) - h_\zeta^*(x)\} \end{aligned}$$

□

B Derivatives

The proof of Part (ii) of Theorem 3.4 is obtained through a sequence of lemmas. We first obtain an alternative representation for the fixed point equation (43). Evaluating this equation at x° , and recalling that $h_\zeta^*(x^\circ) = 0$ gives,

$$\eta_\zeta^* = \zeta \mathcal{U}(x^\circ) + \Lambda_{h_\zeta^*}(x^\circ) \quad (49)$$

Let \mathbb{I} denote the function on X that is identically equal to 1, and for any function h and $\zeta \in \mathbb{R}$ define a new function on X via

$$\mathcal{F}(\zeta, h) := h - \zeta \mathcal{U} - \Lambda_h + [\zeta \mathcal{U}(x^\circ) + \Lambda_h(x^\circ)] \mathbb{I} \quad (50)$$

The fixed point equation becomes (43) becomes,

$$\mathcal{F}(\zeta, h_\zeta^*) = 0. \quad (51)$$

The proof of Theorem 3.4 will require the Implicit Function Theorem. The following version of this result is taken from [14, Theorem 11.5].

Proposition B.1 (Implicit Function Theorem). *Suppose that $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ are open, and that $F: A \times B \rightarrow \mathbb{R}^m$ is continuously differentiable. Suppose moreover that there exists $(x^0, y^0) \in A \times B$ for which the following hold: $F(x^0, y^0) = 0$, and the matrix $\partial/\partial y F(x^0, y^0)$ has rank m .*

Then, there is a ball $O \subset A$ about x^0 and a continuously differentiable function $g: O \rightarrow B$ such the equation $F(x, y) = 0$ is uniquely determined by $y = g(x)$, for each $x \in O$. \square

To apply Prop. B.1, we take $F = \mathcal{F}$ and $(x, y) = (\zeta, h)$, so that $n = 1$ and $m = d$. We apply the result to any $(\zeta_0, h_{\zeta_0}^*)$ to establish that the mapping $\zeta \rightarrow h_\zeta^*$ is C^1 .

For this we require a representation for the derivative of \mathcal{F} with respect to the variable h . The derivative is represented as a $d \times d$ matrix, defined so that for any function $g: \mathsf{X} \rightarrow \mathsf{X}$,

$$\mathcal{F}(\zeta, h + \varepsilon g)|_x = \mathcal{F}(\zeta, h)|_x + \varepsilon \sum_{x' \in \mathsf{X}} \frac{\partial}{\partial h} \mathcal{F}(\zeta, h)|_{x, x'} g(x') + o(\varepsilon)$$

The following follows from (50) and elementary calculus:

Lemma B.2. *The function \mathcal{F} is continuously differentiable in (ζ, h) . The partial derivative with respect to the second variable is,*

$$\frac{\partial}{\partial h} \mathcal{F}(\zeta, h) = I - P_h + \mathbb{I} \otimes \nu,$$

in which P_h is the transition matrix defined in (38), and $\mathbb{I} \otimes \nu$ represents a $d \times d$ matrix with each row equal to ν , and with $\nu(x) = P_h(x^\circ, x)$, $x \in \mathsf{X}$. \square

Invertibility of the derivative with respect to h is obtained in the following:

Lemma B.3. *The following inverse exists as a power series,*

$$Z_h = [I - P_h + \mathbb{I} \otimes \nu]^{-1} = \sum_{n=0}^{\infty} (P_h - \mathbb{I} \otimes \nu)^n$$

in which ν is defined in Lemma B.2. Moreover, νZ_h is the unique invariant pmf for P_h .

Proof. It is easily established by induction that for each $n \geq 1$,

$$(P_h - \mathbb{I} \otimes \nu)^n = P_h^n - \mathbb{I} \otimes \nu_n,$$

where $\nu_n = \nu P_h^{n-1}$. Recall that P_h is irreducible and aperiodic since this is true for P_0 . Consequently, as $n \rightarrow \infty$ we have $\nu_n \rightarrow \pi_h$ and $P_h^n \rightarrow \mathbb{I} \otimes \pi_h$, where π_h is invariant for P_h . The convergence is geometrically fast, which establishes the desired inverse formula.

From the foregoing we have $(P_h - \mathbb{I} \otimes \nu)^n \mathbb{I} = P_h^n \mathbb{I} - \mathbb{I} = 0$ for $n \geq 1$, which implies that $Z_h \mathbb{I} = \mathbb{I}$. From this we obtain,

$$Z_h P_h - \mathbb{I} \otimes \nu = Z_h (P_h - \mathbb{I} \otimes \nu) = Z_h - I$$

Multiplying each side by ν gives $\nu Z_h P_h = \nu Z_h$, so that $\mu_h := \nu Z_h$ is invariant. We also have $\mu_h(\mathbf{X}) = \nu Z_h \mathbb{I} = \nu(\mathbf{X}) = 1$, where we have used again the identity $Z_h \mathbb{I} = \mathbb{I}$. Hence $\mu_h = \pi_h$ as claimed. \square

Since h_ζ^* is continuously differentiable in ζ , it follows from (43) that the same is true for η_ζ^* . The following result provides a representation. The formula for $\frac{d}{d\zeta} \eta_\zeta^*$ could be anticipated from Prop. 2.1.

Lemma B.4. *For each ζ we have,*

$$\frac{d}{d\zeta} \Lambda_{h_\zeta^*} = \check{P}_\zeta H_\zeta^*(x)$$

where $H_\zeta^* = \frac{d}{d\zeta} h_\zeta^*$, and

$$\frac{d}{d\zeta} \eta_\zeta^* = \tilde{\pi}_\zeta(\mathcal{U})$$

Proof. The first result holds by the definition of Λ_h and H_ζ^* . To obtain the second identity, we differentiate each side of (43) to obtain Poisson's equation (44). On taking the mean of each side of (43) with respect to $\tilde{\pi}_\zeta$, and using invariance $\tilde{\pi}_\zeta \check{P}_\zeta = \tilde{\pi}_\zeta$, we obtain,

$$\tilde{\pi}_\zeta(\mathcal{U}) + \tilde{\pi}_\zeta(H_\zeta^*) = \tilde{\pi}_\zeta(H_\zeta^*) + \tilde{\pi}_\zeta\left(\frac{d}{d\zeta} \eta_\zeta^*\right).$$

\square

Proof of Theorem 3.4. Part (i) is contained in Prop. A.2.

Part (ii): Combining Lemma B.2 and Lemma B.3 we see that the conclusions of Prop. B.1 hold for each pair $(\zeta_0, h_{\zeta_0}^*)$. This shows that h_ζ^* is a continuously differentiable function of ζ , and hence \check{P}_ζ is also continuously differentiable. To see that $\tilde{\pi}_\zeta$ is continuously differentiable, apply the representation in Lemma B.3. \square

C Optimal ODE solution

We now prove Theorem 3.5.

The boundary condition is immediate from the assumptions: h_0^* is a constant, since $\check{P}_\zeta = P_0$. Under the assumption that $h_\zeta^*(x^\circ) = 0$ for each ζ , it follows that $h_0^*(x) = 0$ for each x . It remains to show that the relative value function solves the ODE,

$$\frac{d}{d\zeta} h_\zeta^* = \mathcal{H}^\circ(P_\zeta),$$

with \mathcal{H}° defined in (39).

On differentiating each side of (43) we obtain,

$$\mathcal{U}(x) + \frac{d}{d\zeta} \Lambda_{h_\zeta^*}(x) = \frac{d}{d\zeta} h_\zeta^*(x) + \frac{d}{d\zeta} \eta_\zeta^*$$

Based on the definition (34),

$$\Lambda_{h_\zeta^*}(x) = \log\left(\sum_{x'_u} R_0(x, x'_u) \exp(h_\zeta^*(x'_u | x))\right)$$

it follows that $\frac{d}{d\zeta} \Lambda_{h_\zeta^*}(x) =$

$$\left(\sum_{x'_u} R_0(x, x'_u) \exp(h_\zeta^*(x'_u | x)) \right)^{-1} \sum_{x'_u} R_0(x, x'_u) \exp(h_\zeta^*(x'_u | x)) \frac{d}{d\zeta} h_\zeta^*(x'_u | x)$$

The equation simplifies as follows:

$$\begin{aligned} \frac{d}{d\zeta} \Lambda_{h_\zeta^*}(x) &= \sum_{x'_u} \check{R}_\zeta(x, x'_u) \frac{d}{d\zeta} h_\zeta^*(x'_u | x) \\ &= \sum_{x'_u} \check{R}_\zeta(x, x'_u) \frac{d}{d\zeta} \left(\sum_{x'_n} Q_0(x, x'_n) h_\zeta^*(x'_u, x'_n) \right) \\ &= \sum_x \check{P}_\zeta(x, x') \frac{d}{d\zeta} h_\zeta^*(x') \end{aligned}$$

Let $H = \frac{d}{d\zeta} h_\zeta^*(x)$ and $\gamma = \frac{d}{d\zeta} \eta_\zeta^*$. From the foregoing we obtain,

$$\mathcal{U}(x) + \sum_{x'} \check{P}_\zeta(x, x') H(x') = H(x) + \gamma$$

This is Poisson's equation, with $\gamma = \check{\pi}_\zeta(\mathcal{U})$. This shows that h_ζ^* is the solution to the ODE defined in Theorem 3.5, establishing (i) and (ii) of the theorem.

References

- [1] P. BAROOAH, A. BUŠIĆ, AND S. MEYN, *Spectral decomposition of demand-side flexibility for reliable ancillary services in a smart grid*, in Proc. 48th Annual Hawaii International Conference on System Sciences (HICSS), Kauai, Hawaii, 2015, pp. 2700–2709, doi:10.1109/HICSS.2015.325.
- [2] D. BERTSEKAS AND S. SHREVE, *Stochastic Optimal Control: The Discrete-Time Case*, Athena Scientific, 1996.
- [3] D. BERTSEKAS AND J. N. TSITSIKLIS, *Neuro-Dynamic Programming*, Atena Scientific, Cambridge, Mass, 1996.
- [4] V. S. BORKAR, *Convex analytic methods in Markov decision processes*, in Handbook of Markov decision processes, vol. 40 of Internat. Ser. Oper. Res. Management Sci., Kluwer Acad. Publ., Boston, MA, 2002, pp. 347–375.
- [5] S. BOYD AND L. VANDENBERGHE, *Convex Optimization*, Cambridge University Press, New York, 1st ed., 2004.
- [6] R. BROCKETT, *Optimal control of observable continuous time Markov chains*, in IEEE Conference on Decision and Control, Dec 2008, pp. 4269–4274, doi:10.1109/CDC.2008.4738725.
- [7] A. BUŠIĆ AND S. MEYN, *Passive dynamics in mean field control*, in Proc. 53rd IEEE Conference on Decision and Control, Dec 2014, pp. 2716–2721, doi:10.1109/CDC.2014.7039805.
- [8] A. BUŠIĆ AND S. MEYN, *Distributed randomized control for demand dispatch*. arXiv:1603.05966v1 – and to appear, IEEE Conference on Decision and Control, March 2016.

- [9] A. DEMBO AND O. ZEITOUNI, *Large Deviations Techniques And Applications*, Springer-Verlag, New York, second ed., 1998.
- [10] P. GUAN, M. RAGINSKY, AND R. WILLETT, *Online Markov decision processes with Kullback-Leibler control cost*, IEEE Trans. Automat. Control, 59 (2014), pp. 1423–1438, doi:10.1109/TAC.2014.2301558.
- [11] M. KÁRNÝ, *Towards fully probabilistic control design*, Automatica, 32 (1996), pp. 1719–1722.
- [12] I. KONTOYIANNIS AND S. P. MEYN, *Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes*, Electron. J. Probab., 10 (2005), pp. 61–123 (electronic).
- [13] S. KULLBACK, *Certain inequalities in information theory and the Cramer-Rao inequality*, Ann. Math. Statist., 25 (1954), pp. 745–751.
- [14] L. H. LOOMIS AND S. STERNBERG, *Advanced calculus*, Addison-Wesley Reading, MA, 1968.
- [15] A. S. MANNE, *Linear programming and sequential decisions*, Management Sci., 6 (1960), pp. 259–267.
- [16] S. MEYN, P. BAROOAH, A. BUŠIĆ, Y. CHEN, AND J. EHREN, *Ancillary service to the grid using intelligent deferrable loads*, IEEE Trans. Automat. Control, 60 (2015), pp. 2847–2862, doi:10.1109/TAC.2015.2414772.
- [17] S. P. MEYN AND R. L. TWEEDIE, *Markov chains and stochastic stability*, Cambridge University Press, Cambridge, second ed., 2009. Published in the Cambridge Mathematical Library. 1993 edition online.
- [18] M. L. PUTERMAN, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, 2014.
- [19] E. SHAFIEEPOORFARD, M. RAGINSKY, AND S. MEYN, *Rational inattention in controlled Markov processes*, in American Control Conference (ACC), 2013, June 2013, pp. 6790–6797, doi:10.1109/ACC.2013.6580906.
- [20] E. SHAFIEEPOORFARD, M. RAGINSKY, AND S. P. MEYN, *Rationally inattentive control of Markov processes*, SIAM J. Control Optim., 54 (2016), pp. 987–1016, doi:10.1137/15M1008476, http://dx.doi.org/10.1137/15M1008476, arXiv:http://dx.doi.org/10.1137/15M1008476.
- [21] C. A. SIMS, *Implications of rational inattention*, Journal of monetary Economics, 50 (2003), pp. 665–690.
- [22] C. A. SIMS, *Rational inattention: Beyond the linear-quadratic case*, The American economic review, (2006), pp. 158–163.
- [23] R. SUTTON AND A. BARTO, *Reinforcement Learning: An Introduction*, MIT Press. On-line edition at <http://www.cs.ualberta.ca/~sutton/book/the-book.html>, Cambridge, MA, 1998.
- [24] E. TODOROV, *Linearly-solvable Markov decision problems*, in Advances in Neural Information Processing Systems 19, B. Schölkopf, J. Platt, and T. Hoffman, eds., MIT Press, Cambridge, MA, 2007, pp. 1369–1376.